# 对 Boston 数据集的降维分析

CS245　数据科学基础　陆朝俊
叶泽林 515030910468

## 1　问题描述

在数据预处理中，数据约简是一个重要的步骤，数据约简技术可以得到数据集的约简表示，即缩小数据容量但保持了原始数据的大多数信息，使得之后的分析更加高效，而分析结果与未约简的结果几乎相同。

在数据量和数据复杂性日益增多情况下，数据约简更是数据预处理中不可或缺的关键一环。这次作业中，我将使用数据约简技术（以 PCA 为主）对 Boston 数据集进行降维分析。

## 2　解决方案[1]

### 2.1　数据集获取及读入

Boston 数据集全称为波士顿房价数据集（Boston House Price Dataset），给定房屋及其相邻房屋的详细信息进行房价预测，是一个针对回归问题的数据集。该数据集包含 506 个实例，每个实例拥有 13 个特征及一个回归目标（房价），具体的数据集特征情况可参见附录 A.1。

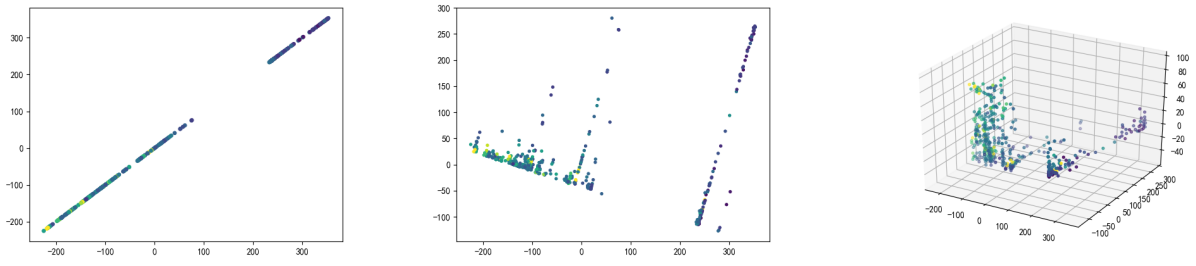
Boston 数据集已集成在 Python 的 scikit-learn 模块下，安装好该模块后只需运行以下代码即可读取数据集：

```
1 from sklearn import datasets
2 boston = datasets.load_ boston()
```

### 2.2　降维分析

降维分析中最常用的就是主成分分析（PCA）算法，本次作业中也将采用 PCA 算法进行降维分析。PCA 的主要思想是：找出能反映最大偏差的特征线性组合（即主成分），构成新特征空间。

## 3　结果展示



(a) 主成分为 1　(b) 主成分为 2　(c) 主成分为 3

图 1: PCA 效果可视化（图中不同点的颜色代表不同的回归目标，即房价，颜色越接近则表示数值越接近）

[1]本次作业的所有代码实现可参见附录 A.2

表 1: Boston 数据集 PCA 降维效果比较

| 主成分个数 | 主成分方差占比之和 |
|---|---|
| 1 | 0.80581 |
| 2 | 0.96887 |
| 3 | 0.99021 |
| 4 | 0.99717 |
| 5 | 0.99848 |
| 6 | 0.99921 |
| 7 | 0.99963 |
| 8 | 0.99988 |
| 9 | 0.99996 |
| 10 | 0.99999 |
| 11 | 0.99999 |
| 12 | 0.99999 |
| 13 | 1.00000 |

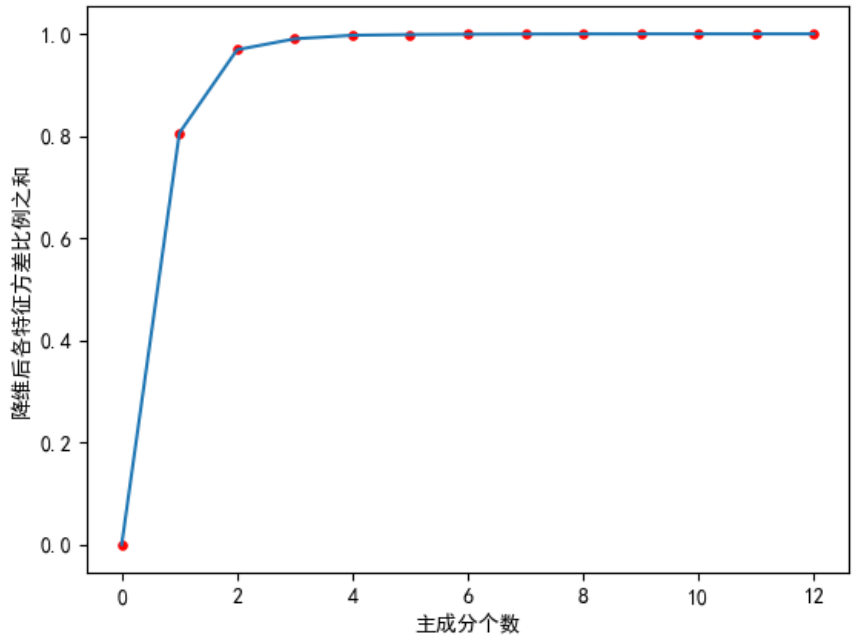



图 2: PCA 对于 Boston 数据集的降维效果

# A 附录

## A.1 Boston 数据集特征信息

表 2: Boston 数据集特征信息

| 编号 | 特征名 | 特征含义 |
|---|---|---|
| 1 | CRIM | 城镇人均犯罪率 |
| 2 | ZN | 住宅用地超过 25000 sq.ft. 的比例 |
| 3 | INDUS | 城镇非零售商用土地的比例 |
| 4 | CHAS | 查理斯河空变量（如果边界是河流，则为 1；否则为 0） |
| 5 | NOX | 一氧化氮浓度 |
| 6 | RM | 住宅平均房间数 |
| 7 | AGE | 1940 年之前建成的自用房屋比例 |
| 8 | DIS | 到波士顿五个中心区域的加权距离 |
| 9 | RAD | 辐射性公路的接近指数 |
| 10 | TAX | 每 10000 美元的全值财产税率 |
| 11 | PTRATIO | 城镇师生比例 |
| 12 | B | $1000(Bk-0.63)^2$，其中 Bk 指代城镇中黑人的比例 |
| 13 | LSTAT | 人口中地位低下者的比例 |

## A.2 主要代码